Noisy Stereotypes[*][†]

March 6, 2020 Draft – please do not cite or circulate without authors' permission

Abstract

Many philosophers have argued that statistical evidence regarding group characteristics can create normative conflicts between the requirements of epistemic rationality and our moral obligations to each other. In a recent paper, Johnson-King and Babic argue that such conflicts can usually be avoided: what ordinary morality requires, they argue, epistemic rationality permits. However, we show that as data gets large, Johnson-King and Babic's approach becomes implausible. More constructively, we develop an alternative model of reasoning about stereotypes under which one can indeed avoid normative conflicts even in a big data world, but only if such data contain some noise.

[*] Author correspondence.

[†] Acknowledgements.

1 Introduction

In a world characterized by socioeconomic and other inequalities, some stereotypes will be statistically sound. In those cases, many philosophers have argued, epistemic rationality can come apart from our moral obligations to each other (e.g., Gendler, 2011; Basu, 2018; Basu and Schroeder, 2018). For example, Tamar Gendler puts the point as follows:

> As long as there's a differential crime rate between racial groups, a perfectly rational decision maker will manifest different behaviors, explicit and implicit, toward members of different races. This is a profound cost: *living in a society structured by race appears to make it impossible to be both rational and equitable* (Gendler, 2011, p. 57, emphasis added).

Gendler calls this the Sad Conclusion.

In a recent paper, Johnson-King and Babic (2019) argue that such normative conflicts – i.e., those giving rise to Gendler's Sad Conclusion – can usually be avoided: what ordinary morality demands, they argue, epistemic rationality typically permits. To develop their argument, they rely on the notion of minimizing epistemic risk, developed in Babic (2019), as a principle for identifying an appropriate prior in the absence of information. In this project, however, we explain that as a dataset gets large, Johnson-King and Babic's approach to avoiding normative conflicts becomes less persuasive. More constructively, we build on their project and develop an alternative model of reasoning about stereotypes under which one can indeed avoid normative conflicts, even in a big data world, but only if such data contain some noise. In doing so, we also articulate a model of rational belief updating in response to learning experiences characterized by large but noisy samples.

The paper proceeds as follows. We first explain the basic notion of epistemic risk and describe the argument in Johnson-King and Babic (2019). A key step in their argument is that different attitudes to epistemic risk license different priors in the absence of other information. And in most cases giving rise to normative conflicts, there will exist an epistemically permissible prior which cautions an agent from adopting stereotype reinforcing credences (for example: a prior which cautions against adopting a high credence that some racial groups are more likely to commit certain crimes, in Gendler's example). We then articulate our challenge to this argument: with large datasets, tweaking the priors only goes so far – the likelihood dominates inferences and as a result normative conflicts will inevitably reemerge. Finally, and most importantly, we develop a model of statistical inference under noise and use it to explain how such normative conflicts can be avoided still, when the data is not perfect. Our model leaves room for identifying true population differences where they really exist. Most differences giving rise to normative conflicts, or the Sad Conclusion, as Gendler puts it, are the result of various types of bias.

2 Background and framework

We develop the argument to follow within the general framework of epistemic utility theory.[1] In particular, we assume that an epistemically rational agent should adopt credences in a way that minimizes expected inaccuracy, where inaccuracy is measured by an appropriate scoring rule. To keep things concrete, consider a simple example of a weather forecaster predicting the probability of rain tomorrow. Higher probabilities are "better" if it rains, lower probabilities are "better" if it does not rain. We then have the following setup:[2] A measurable space Ω of states, a σ-algebra $\mathcal{F}$ of events, and a probability measure $\mathbb{P}$ reflecting the agent's beliefs. The triple $(\Omega, \mathcal{F}, \mathbb{P})$ constitutes our subjective probability space. The truth value of $A \in \mathcal{F}$ is represented by

$$\mathbb{I}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Since the truth is unknown, $\mathbb{P}[A]$ should be one's best estimate of $\mathbb{I}_A$, where best is defined in terms of (expected) inaccuracy.

Given $(\Omega, \mathcal{F}, \mathbb{P})$ and $A \in \mathcal{F}$, let $s : \mathbb{P}[A] \times \mathbb{I}_A \to \mathbb{R}$ be a measure of the inaccuracy of credences given by $\mathbb{P}$ relative to A. This is a scoring rule or loss function for evaluating the accuracy of credences. For example, $s(\mathbb{P}[A], \mathbb{I}_A) = (\mathbb{P}[A] - \mathbb{I}_A)^2$ corresponds to the well-known Brier (1950) score (expressed here so that a lower score is better). Four properties of scoring rules will be relevant as we develop our argument. They are as follows. Let $\mathbb{P}[A] = p$. A scoring rule is

(1) **Monotonic**, if $s(p, 1)$ is decreasing in p and $s(p, 0)$ is increasing in p.

(2) **Continuous**, if $s(p, \mathbb{I}_A)$ is a continuous function of p.

(3) **Strictly Proper**, if $\mathbb{E}_p[s(p, \mathbb{I}_A)] < \mathbb{E}_p[s(q, \mathbb{I}_A)] \;\; \forall\, q \neq p$.[3]

These properties, together with some modest decision-theoretic norms, commit us to **Probabilism** – the thesis that subjective credences should conform to the probability axioms – and **Conditionalization** – the thesis that upon receiving new information (in ordinary circumstances), one should update credences by Bayes' Rule. Joyce (2009)

[1] See, e.g., Joyce (1998), Greaves and Wallace (2006), Joyce (2009), Pettigrew (2016), Huttegger (2017).

[2] The notation we use generally follows Huttegger (2013, 2017).

[3] If the score is defined so that a higher score is better, the inequality is reversed.

shows that credences which are not probabilities are strongly accuracy-dominated for every scoring rule satisfying (1)-(3). Greaves and Wallace (2006) and Huttegger (2013) show that updating beliefs by Bayesian conditioning minimizes expected inaccuracy for every accuracy measure satisfying (1)-(3). This, in a nutshell, is what we take to be a floor on epistemic rationality.

One more property will be relevant to explain the notion of epistemic risk, below, but we do not assume it to be a requirement of epistemic rationality. A scoring rule is

(*) **0-1 Symmetric**, if $s(p,1) = s(1-p,0) \ \forall\, p \in [0,1]$.

One may or may not evaluate accuracy with a 0-1 Symmetric scoring rule.

3 Epistemic Risk

To keep things simple, consider again a dichotomous event, A. For example, $A :=$ 'Alice will receive tenure in 2020'. Let

$$\Omega = \{A, A^c\},$$
$$\mathcal{F} : 2^{\Omega}, \text{ and}$$
$$\mathbb{P}[A] = p.$$

The core idea in Babic (2019) is that there are two ways of being inaccurate about A. We may increase our credence in A when A is false. This is a case of increasing inaccuracy in the false positive direction. Or we may decrease our credence in A when A is true. This is a case of increasing inaccuracy in the false negative direction. For scoring rules that are 0-1 symmetric, a unit increase in inaccuracy in either direction should be treated equally. But there is no reason why this should be so. For instance, we may care more about falsely assuming that Alice will not get tenure if she in fact turns out to get it than we do about falsely assuming she will get tenure if she in fact does not. It depends on what is at stake. However, the way one adjudicates the relative costs of approaching inaccuracy in either direction determines the scoring rule they deem appropriate.

Let $r : [0,1] \to \mathbb{R}$ be a continuous, twice-differentiable, strictly concave function of p which associates each credence with some amount of epistemic risk. For now, $r(p)$ is just an arbitrary function satisfying these constraints. However, the constraints are such that the symmetry and curvature of $r(p)$ will reflect the agent's attitudes to error and from it we can derive the scoring rule she deems appropriate. For example, consider the following two possible epistemic risk functions and their interpretations with respect to the agent's attitudes to error.

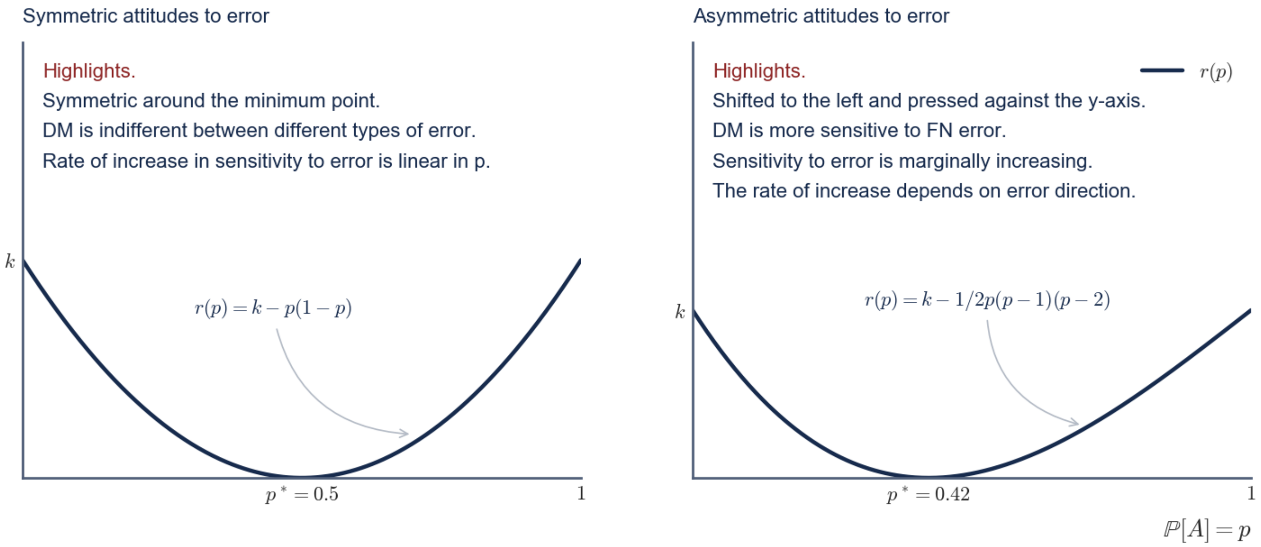

Figure 1: Two possible epistemic risk functions and their associated interpretations.

We can characterize, from $r(p)$, the class of scoring rules satisfying (1)-(3) through a pair of ordinary differential equations. This is a consequence of Theorem 1 from Babic (2019), together with Savage (1971). If for $p \in [0, 1]$,

$$\begin{aligned} &(1)\ r(p) \geq 0 \\ &(2)\ r''(p) \in [0, \infty) \end{aligned}$$

then, letting $k = \max\ r(p)$,

$$\begin{aligned} s(p, 1) &= k - r(p) - (1-p)r'(p) \\ s(p, 0) &= k - r(p) + pr'(p) \end{aligned} \tag{1}$$

are strictly proper scoring rules. Figure 2 should make this more intuitive. On the left-side we have the symmetric epistemic risk function from the left-side of Figure 1, and on the right-side we have the asymmetric epistemic risk function from the right-side of Figure 1, together with the scoring rules derived from each.

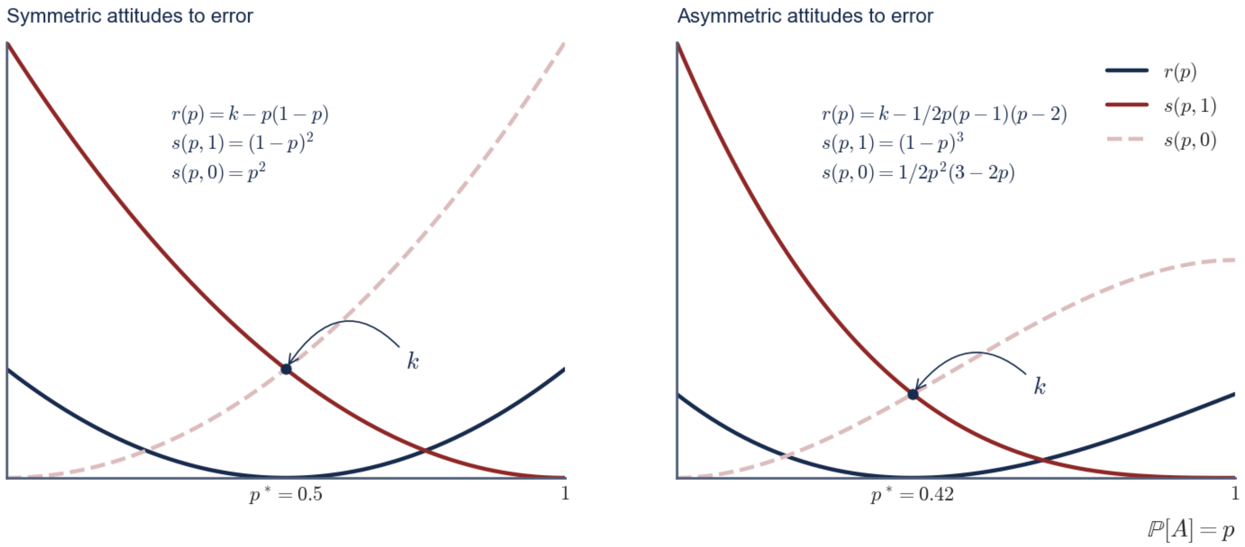

Figure 2: Two possible epistemic risk functions and their associated scoring rules.

Importantly for us, if p^* is centered at 0.5 the scoring rule will treat increases in inaccuracy in the false positive error direction as bad as increases in inaccuracy in the false negative error direction. This is a symmetric epistemic risk function. If $p^* > 0.5$, false positive increases in inaccuracy are more costly. If $p^* < 0.5$, the reverse is true. In the next section, we will explain how Babic and Johnson-King use this framework to argue that Gendler's Sad Conclusion will rarely, if ever arise.

4 Normative Conflicts

The epistemic risk approach provides a general recipe for identifying priors in the absence of information. We ask, first, how do we care about the relevant errors; and, second, how much epistemic risk are we willing to assume? Consider again as an example the pair of functions we have been using throughout, and assume that our agents seek to minimize epistemic risk. Then, first, their prior point estimate for p should be p^*.

More interesting, however, is the sort of case considered in Johnson-King and Babic (2019) ("JKB" henceforth). Many philosophers have pointed out that in a society characterized by substantial socioeconomic inequality, we will often be in a position where we learn about uncomfortable base rates regarding vulnerable or historically discriminated groups. For example, many personal traits – such as prior criminal convictions, socioeconomic status, and some medical conditions – are subject to social stigma. The prevalance of these traits is undoubtedly unequally distributed across different demographic groups (racial, gender, ethnic, class, etc.). Gendler's Sad Conclusion refers to the notion that in an unequal society we will often learn about the unequal distribution of stigmatized traits along sensitive demographic lines. We will then, it seems, be forced by the hand of epistemic rationality to formulate beliefs about vulnerable groups that strike most people as immoral (see e.g., Basu and Schroeder, 2018).

Situations that give rise to the Sad Conclusion (or as JKB and others call it, to Normative Conflicts) are ordinarily of the following form: we learn some statistical information pertaining to the prevalence of a certain trait (job performance, criminality, creditworthiness, etc.) within a group (gender, racial, ethnic, etc.). We then meet an individual who belongs to the group, and we have to formulate a credence about how likely the individual is to possess the trait. JKB use the following example.

> **Gender Bias Study**. One morning, you read a report about a study on gender discrepancies in academic employment. The study surveyed 500 men and 500 women employed in universities. They found that only 30% of the women were employed in faculty positions, while the other 70% were administrative assistants. For men, the proportions were reversed. Before learning this, you had no prior relevant information. The study was otherwise legitimate. You then meet Mary. Mary tells you that she works in a university. What should be your credence that Mary is a faculty member?

Many epistemologists would argue that epistemic rationality requires one to believe it is 70% likely that Mary is an administrative assistant. Call this the Naive Answer. JKB argue against the Naive Answer by explaining that what epistemic rationality requires ultimately depends on one's antecedent attitudes to epistemic risk. The question we have to answer here is how to formulate a predictive inference about Mary.

JKB follow the standard Bayesian answer to this question, which we develop and refine here, as it will be important to understanding our revised model later. Let θ be the (unknown) proportion of women in academia who are faculty. Suppose that in a sample of n women in academia, we observe x women that are faculty. Then x follows a Bernoulli process with parameter θ and the likelihood function is given by

$$\ell(x|\theta, n) = \theta^x(1-\theta)^{n-x} \tag{2}$$

In the Bayesian approach, we need to identify our prior beliefs regarding θ. A beta distribution being a relatively flexible distribution can approximate a wide variety of information regarding a Bernoulli process and is commonly used in Bayesian models involving proportions (e.g., Lindley and Phillips, 1976). Let $f(\theta)$ be the prior probability density for θ, where

$$f(\theta) = f_\beta(\theta|a_\theta, b_\theta) = \theta^{a_\theta - 1}(1-\theta)^{b_\theta - 1}/B(a_\theta, b_\theta) \tag{3}$$

is a beta density function with $B(a_\theta, b_\theta) = \Gamma(a_\theta)\Gamma(b_\theta)/\Gamma(a_\theta + b_\theta)$. The mean and variance of a beta distribution are given by $\mathrm{E}[\theta] = a_\theta/(a_\theta + b_\theta)$ and $\mathrm{Var}(\theta) = a_\theta b_\theta/(a_\theta + b_\theta)^2(a_\theta + b_\theta + 1)$. As a_θ becomes larger the distribution moves towards the right, whereas an increase in b_θ moves the distribution towards the left. When $a_\theta = b_\theta$, the distribution is symmetric around 0.5. If both a_θ and b_θ increase then the distribution begins to narrow. The parameters a_θ and b_θ can also be interpreted as "pseudo" observations upon which the prior beliefs are based. For instance, $a_\theta = 7$ and $b_\theta = 3$ would be equivalent to having observed 7 faculty and 3 non-faculty out of 10 women in academia. These properties will be important to our argument; other features of the beta distribution can be easily found in any standard Bayesian textbook, such as Gelman et al. (2013).

With the likelihood in (2) and the prior in (3), the posterior density for θ is given by

$$f(\theta|x, n) = f_\beta(\theta|a_\theta + x, b_\theta + n - x) \propto \theta^{a_\theta + x - 1}(1-\theta)^{b_\theta + (n-x) - 1} \tag{4}$$

Note that the posterior distribution is of the same form as the prior distribution. This is because a beta distribution is *conjugate* to the Bernoulli process. This means that if we start with a beta prior for θ, update that via Bayes' Rule with data from a Bernoulli process, our posterior will be also be a beta but with updated parameters. Such a model lends itself to an intuitive interpretation that is helpful to keep in mind. The posterior beta distribution for θ after seeing the data is given by simply adding the actual observations (in the sample data) and the corresponding pseudo observations represented in the prior distribution. For example, suppose our prior on θ is a beta density with parameters $(a_\theta = 7, b_\theta = 3)$, and in the sample data we observe 4 out of 10 women in academia who are faculty, our posterior for θ would be a beta density with parameters (7+4, 3+6).

But in order to formulate a credence about Mary, we need more than the posterior distribution. Let $\widetilde{X} \in \{0, 1\}$ be an additional outcome that has yet to be observed (i.e., Mary). The distribution of $\widetilde{X}$ given x is called the predictive distribution, and is of the following form:

$$\begin{aligned} P(\widetilde{X} = 1|x) &= \int_0^1 P(\widetilde{Y} = 1|\theta, x) f(\theta|x) d\theta \\ &= \int_0^1 \theta f(\theta|x) d\theta = E(\theta|x) = \frac{a_\theta + x}{a_\theta + b_\theta + n}. \end{aligned} \tag{5}$$

The expression in (5) is consistent with what Huttegger (2017) calls the Generalized Rule of Succession. Huttegger (2017), following Zabell (2005), Carnap (1950) and Johnson

(1924), shows that this form of the predictive probability follows from some very modest assumptions about the structure of the data-generating distribution, which are satisfied by the Bernoulli process we have here. The question for us is: which values should we assign to a_θ and b_θ? Equivalently: which prior for θ should we adopt in the absence of information?

First, the Naive Answer requires that $a_\theta = b_\theta = 0$. This would result in an improper (and as a result, arbitrarily incoherent) prior. Second, following Laplace's Rule of Succession, we could set $a_\theta = b_\theta = 1$. This is equivalent to a uniform prior for θ. And third, JKB follow Huttegger's Generalized Rule of Succession, but they use attitudes to epistemic risk in order to identify appropriate values for a_θ and b_θ. If, for example, we seek to minimize epistemic risk in the absence of information, then we must use a prior for θ which is such that $\mathrm{E}[\theta] = \arg\max\ r(P)$. Any prior satisfying this constraint will be permitted given one's epistemic risk function. This is again easier to explain by illustration.

Recall that in our notation $\arg\max\ r(P) = p^*$ and $r(p^*) = k$. Figure 3, below, depicts various permissible distributions for θ, given some k, which is determined by the agent's attitudes to epistemic risk. The dot in the left panel corresponds to the same k from the left panels in the preceding figures, and the dot in the right panel corresponds to the same k from the right panels in the preceding figures. However, we have removed the epistemic risk function and scoring rules and added the appropriate prior distributions, given those epistemic risk functions and scoring rules.

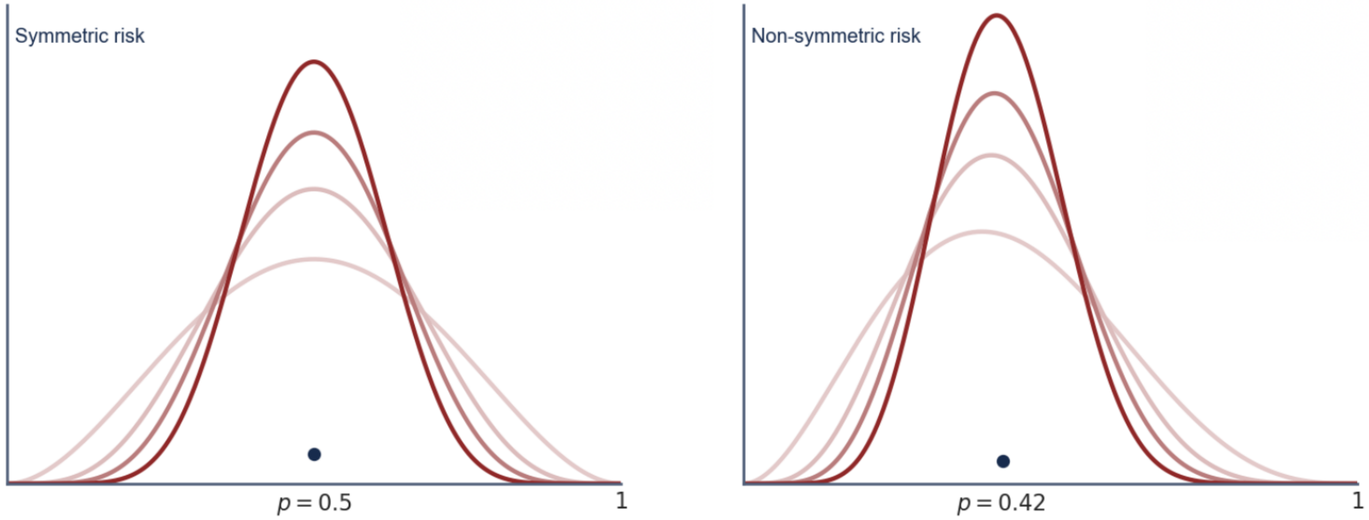

Figure 3: Permissible prior distributions for θ, given an epistemic risk function and a desire to minimize epistemic risk.

For example, suppose we find it more costly to increase confidence in the claim that Mary is not faculty when she is than we do in the claim that Mary is faculty when she is not. Then, treating the relevant proposition as 'Mary is not-faculty', we should choose one of the distributions from the right-panel of Figure 5, because any of these err in the direction of high credence in the claim that Mary is faculty (if we treat the relevant proposition as 'Mary is faculty' we would pick a distribution where $\mathrm{E}[\theta] > 0.5$). Each distribution in Figure 5 corresponds to particular values for a_θ and b_θ in Huttegger's Generalized Rule of Succession. Thus, identifying a distribution for θ is equivalent to specifying how one will apply the Generalized Rule of Succession. It is now a short step to see how JKB seek to avoid normative conflicts. The general scheme is this, which we call the JKB Approach.

> **JKB Approach.** If one wants to end up with at most a middling (0.5) probability for $\widetilde{X}$, then given a sample mean x/n one must start with a Beta prior for θ which is such that $a_\theta \leq n - x$ and $b_\theta = n - a_\theta$.

Plugging these values into Huttegger's Generalized Rule of Succession, one will guarantee that the updated credence for the claim that Mary is not faculty will not be above 0.5, regardless of how strong the data in the Gender Bias Study is. In their paper, JKB consider a case where the data indicates that of 1000 women, 300 are faculty. They show that a rational epistemic agent can avoid normative conflicts by adopting a Beta prior with $a_\theta = 700$, and $b_\theta = 300$.

5 Stereotypes Under Big Data

The JKB Solution explains many cases where normative conflicts might have arisen in the past. Like the Gender Bias Study, if all you have to go on is a newspaper description of some isolated report, perhaps it's worth being extra careful as you apply this to Mary. But it is important to be clear about the conceptual approach here: we avoid a certain kind of prediction by loading our prior in a way that hedges against a particular conclusion. While this can be appropriate, to an extent, as the sample size increases we are vulnerable to the objection that we have buried our head in the sand, so to speak. Moreover, as the sample size increases, we have to load our prior to such an extent that we become increasingly stubborn and unable to learn. For example, the figure below illustrates how sharp one's prior would have to be in order to avoid the Sad Conclusion using the JKB Approach, for a sample size of $5,000$, $10,000$, and $100,000$. In each case, the sample mean is assumed to remain as in the original hypothetical, i.e., 0.7.

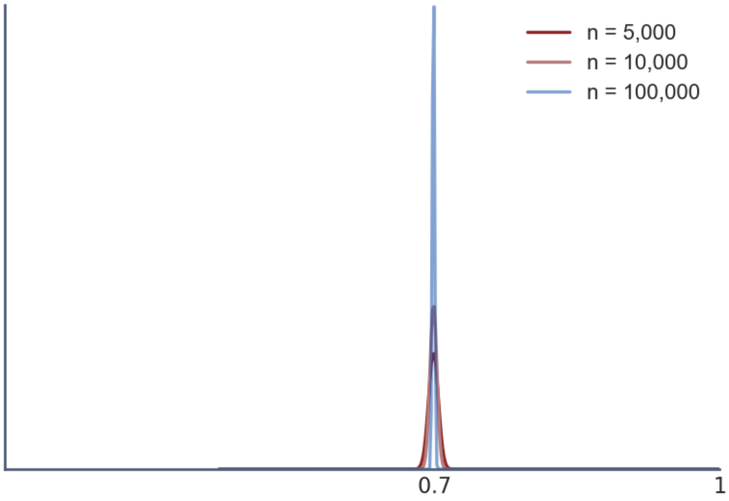

Figure 4: Prior distributions for θ, if one is to avoid the Sad Conclusion, for $n = 5000$, $n = 10000$, $n = 100000$.

Notice that as we get to $n = 100,000$, the prior becomes arbitrarily peaked around the agent's risk-free point, i.e., the point we have called $r(p^*) = k$. At this stage, one's attitudes to epistemic risk are so dominant, that the prior is essentially approaching a Dirac delta function which concentrates all its mass on k. Moreover, as the dataset becomes larger, the argument we will make here becomes increasingly more salient: the more observations we have, the closer one's prior will have to get to an arbitrarily peaked density around the risk-free point in order to avoid normative conflicts. And $100,000$ observations is not especially large. Facebook, Google, and Amazon, for example, collect data on billions of people. There is no hope, in cases like this, of avoiding normative conflicts by tweaking one's priors by reference to underlying attitudes to epistemic risk.

6 Belief Updating in the Face of Noise

With perfect data, the reemergence of the Sad Conclusion is perhaps inevitable. But large datasets are rarely perfect. When we collect statistical evidence about people, such as in the Gender Bias Study, there typically exists some classification error. Insofar as the Gender Bias Study assumes perfect sampling, it is not representative of the types of cases we usually face. Suppose, for example, we survey people about their voting habits, asking whether they will vote Democrat or Republican in the next presidential election. While most people may be honest, some will state that they will vote Republican when they in fact vote Democrat, and others will state they will vote Democrat when they in fact vote Republican. It is also possible that the errors are one-sided – for example, only Democrats misprepresent themselves while Republicans do not, or vice versa – or that the errors occur on both sides but at uneven rates. Noise is the norm rather than the exception when we collect data.

While the voting example does not necessarily involve a pernicious predictive inference, as JKB call them, noise is similarly pervasive in the kinds of cases that lead to the Sad Conclusion. Consider, as another example, the now well-known COMPAS recidivism algorithm (Angwin et al., 2016; Kleinberg et al., 2016). In some US jurisdictions, judges in state courts use an algorithm, called COMPAS, in order to decide whether a defendant in a criminal proceeding should be released on bail or not pending trial. The algorithm uses demographic data about defendants in order to make a prediction about their future behavior. The predictions are made in the form of a decile based two-year recidivism risk score.

In the actual data, which has been made available through Freedom of Information Act Requests,[4] and profiled by the magazine Pro Publica, the probability that a randomly

[4] The data can be found here, https://github.com/propublica/compas-analysis/.

selected black defendant is classified as high risk is much higher than the probability that a randomly selected white defendant is classified as high risk (Angwin et al., 2016). This is in part because in the historical record black defendants have on average more prior criminal convictions, and priors are especially probative of future criminal behavior. This is exactly the sort of case that gives rise to the Sad Conclusion. The statistical evidence, on its own, suggests that a black is defendant is more likely to commit a crime if released than a white defendant.

Suppose we had no prior information about criminal behavior across race. When we learn about the COMPAS data, how should we formulate our credences? The Naive Answer would suggest that our credence should correspond to the algorithmic prediction. This leads to the Sad Conclusion. The JKB recipe would suggest we adopt a prior which is extremely sensitive to false positive mistakes with respect to black defendants. This avoids the Sad Conclusion only insofar as the dataset is not too large – in the example of recidivism prediction, however, we have records for 100,000 individuals.

However, consider what else we know about cases like this: we know that we live in a world characterized by racial and socioeconomic inequality, that black individuals are stopped and often falsely targeted at higher rates, that they are more vigorously prosecuted, that they are more likely to be convicted (including for petty and victimless or small drug crimes), that they are more likely to be imprisoned and receive a criminal record, etc. This suggests that in the data we are much more likely to see non-violent black people falsely classified as high risk than we are to see non-violent white people falsely classified as high risk. Indeed, this is what we find if we look through the data. For example, in the original COMPAS data, we can find many examples like the following (to stress: these are real cases):

> **COMPAS Misclassification.**
>
> Kevin McManus is a 53 year old white male with 19 prior criminal convictions. The predicted probability that he will recidivate if released is 50%.
>
> Kiante Slocum is a 21 year old black female with 2 prior criminal convictions. The predicted probability that she will recidivate if released is 80%.

As a result, if we learn about the algorithm's output, before identifying our credence with that output, we need to consider the noise or misclassification rate – the rate at which both white and black defendants are falsely classified as high or low risk. The JKB recipe does not consider this – because it assumes exchangeable data drawn from a Bernoulli distribution.

We will now develop a simple model of belief updating under noise and show how the Sad Conclusion can be avoided, regardless of the size of the dataset, when the data is noisy (in the way that COMPAS data is, for example).

To develop this model, we will use a stylized case like the Gender Bias Study. While our model applies to cases like COMPAS, for our purposes it is unnecessarily complicated. There are many race categories, the output is given in terms of decile scores, and there is the added layer of an algorithmic prediction. Instead, we will stick to a simple case with two categories (male and female), and two outputs (a favorable class and an unfavorable class). But, unlike the Gender Bias Study, we want an example more representative of real life survey sampling, where the possibility of error exists, and where the dataset is large. Consider the following.

> **Recruitment**. A study on gender disparities in performance among investment bankers looked at evaluations of 100,000 men and 100,000 women in junior positions. The evaluations recorded each employee as "partnership worthy" or "not partnership worthy" (Talented and Untalented, for short). The researchers found that only 30% of the women were recorded as talented whereas for men, 70% were recorded as talented. The research is otherwise sound.
>
> Later, you meet Alice, who has applied for a job at your bank. How confident are you that Alice is **actually** Talented, based on data regarding the proportion of women who were recorded, or **deemed**, Talented?

Using the JKB recipe, the probability would be proportional to $p(\theta)f(\mathbf{x}|\theta)$ where θ follows a Beta distribution, with a_θ and b_θ being determined by the agent's attitudes to epistemic risk, as in Figure 3, and each x is a Bernoulli draw (corresponding to each woman being Talented or Not).

In the Gender Bias Study, there were 500 women. In our example, there are 100,000 women. This already makes the JKB recipe inappropriate because of the arbitrarily peaked prior it would require. But the JKB model is not even appropriate for Recruitment, because it ignores the possibility that an actually Talented woman is recorded as Untalented. This is a substantial risk in a world we know to be characterized by gender inequality, in which women face especially large barriers to success, and in which implicit bias regarding gender affects women's performance assessments. Thus, we need a different model for prediction here altogether. The following material will be somewhat technical, so before we proceed let us explain the motivation, which is actually quite intuitive.

Recall that in the beta-binomial model, updating beliefs by Bayes' Rule is equivalent to counting the number of favorable and unfavorable observations. Thus, when there is no noise, as in the Gender Bias Study, to compute the posterior we simply add the number of women faculty to our initial value of a_θ, and the number of women non-faculty to our initial value of b_θ. When there is noise, as in Recruitment, we don't want to do this, because we suspect that the number of women deemed Untalented is inflated. Therefore, given some noise rate, and a sample size, what we want to figure out is the "noise-adjusted" sample size, and update on *that* instead. For example, our model will allow us to make statements like the following: if the sample is 1000 women, and the misclassification rate is 20%, then given a beta-binomial model, the equivalent noise-free sample to plug into one's update is actually 10 women. What will be interesting to observe is how rapidly the noise-adjusted sample size decreases as the misclassification rate increases. This is how we avoid the Sad Conclusion in the context of noisy big data.

To be clear about our notation, let $\theta \in [0, 1]$ be the proportion of women who would be truly successful under ideal conditions, i.e., conditions identical to those of men (or as an alternative interpretation, those who would be promoted absent any socioeconomic, occupational, political etc. gender disparities among men and women). The point is, θ represents women's true Talent rate – it is an unobserved latent variable, much like IQ, EQ, or any other such indicator of aptitude in education or the workforce. Further, let $\lambda \in [0, 1]$ be the misclassification rate of Talented women as Untalented. We assume that the other type of misclassification (Untalented women recorded as Talented) is so small that it is not worth worrying about. Then, the data generating model we have visually looks like this.

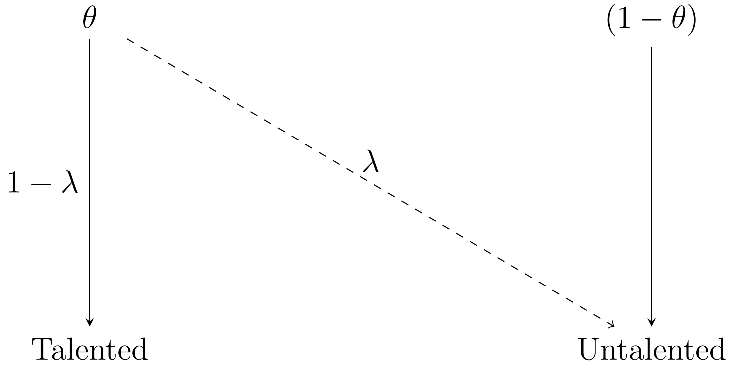

Figure 5: Recruitment with noise level λ

In this model, the probability that a woman is deemed Talented is $\phi = \theta(1-\lambda)$. The probability that a woman is deemed Untalented is $1 - \phi = (1-\theta) + \theta\lambda$. In a sample of n women, let γ be the number of women who are deemed Talented and the remaining $n - \gamma$ as Untalented. Then the data generating process for the recording (as opposed to the actual number) of women as Talented and Untalented is Bernoulli in ϕ and not in θ (the actual proportion of Talented women). The likelihood of the sample is thus of the form

$$\ell(\gamma|n,\theta,\lambda) = \big[\theta(1-\lambda)\big]^{\gamma}\big[(1-\theta)+\theta\lambda\big]^{n-\gamma} \tag{6}$$

The maximum likelihood estimate of (θ, λ) is not unique. The likelihood function is unable to distinguish among all (θ, λ) pairs with the same value of ϕ. As a result of this identification problem, the maximum likelihood estimate of (θ, λ) consists of all (θ, λ) pairs such that $\theta(1-\lambda) = \gamma/n$. For example, if 30% of 100,000 women are observed as Talented, then the likelihood is maximized at an infinite combinations of (θ, λ) including $(\theta = 0.3, \lambda = 0), (\theta = 0.5, \lambda = 0.4), (\theta = 0.8, \lambda = 0.625), (\theta = 1, \lambda = 0.7)$, and many more. In other words, many disparate explanations of the data seem equally compelling. If we assume that $\lambda = 0$ then this model is the same as the JKB noise-free model.

The likelihood function in (6) can be expressed as

$$\ell(\gamma|n,\theta,\lambda) = \sum_{t=0}^{n-\gamma} \binom{n-\gamma}{t} \theta^{n-t}(1-\theta)^{t}\lambda^{n-\gamma-t}(1-\gamma)^{\gamma} \tag{7}$$

Here t can be interpreted as the number of women who are correctly classified as Untalented, or equivalently $n - \gamma - t$ as the number of women who are incorrectly classified as Untalented. Since we do not know or observe t, the likelihood is expressed as a mixture of the $n - \gamma + 1$ likelihoods that could arise with each possible number of misclassifications in the data.

While λ is unknown, one might have some prior beliefs on λ along with those on θ. We might use our background knowledge about social inequality, gender stereotypes in finance, barriers to success for women in business, historical practices of discrimination, and so forth. For instance, *a priori*, $(\theta = 0.7, \lambda = 0.3)$ might be considered more likely than $(\theta = 0.49, \lambda = 0)$ or $(\theta = 0.98, \lambda = 0.5)$, although all three pairs have identical likelihoods for any given data. The Bayesian approach encourages us to use all this prior information, rather than ignoring it and blindly following the data. Such beliefs can be represented in a prior distribution for θ and λ, and then given a sample, the beliefs can be updated in a Bayesian manner.

Bayesian models with unknown misclassifaction rates in dichotomous data have been developed in Winkler and Gaba (1990), Gaba and Winkler (1992), and Gaba (1993). For ease of exposition here, we restrict attention to their special case with a prior density on (θ, λ) that assumes θ and λ are *a priori* independent and is given by

$$\begin{aligned} f(\theta,\lambda) &= f_{\beta}(\theta|a_{\theta},b_{\theta})f_{\beta}(\lambda|a_{\lambda},b_{\lambda}) \\ &\propto \theta^{a_{\theta}-1}(1-\theta)^{b_{\theta}-1}\lambda^{a_{\lambda}-1}(1-\lambda)^{b_{\lambda}-1} \end{aligned} \tag{8}$$

With the prior in (8), and the likelihood in (7), the posterior density is given by

$$\begin{aligned}
&f(\theta, \lambda|\gamma, n) = \sum_{t=0}^{n-\gamma} w_t f(\theta, \lambda|\gamma, n, t),\\
&\text{where}\\
&w_t = a_t / \sum_{t=0}^{n-\gamma} a_t,\\
&a_t = \binom{n-\gamma}{t} \mathrm{B}(a_\theta^*, b_\theta^*)\mathrm{B}(a_\lambda^*, b_\lambda^*),\\
&f(\theta, \lambda|\gamma, n, t) = f_\beta(\theta|a_\theta^*, b_\theta^*) f_\beta(\lambda|a_\lambda^*, b_\lambda^*),\\
&\text{with}\\
&a_\theta^* = a_\theta + n - t\\
&b_\theta^* = b_\theta + t\\
&a_\lambda^* = a_\lambda + n - \gamma - t, \text{and}\\
&b_\lambda^* = b_\lambda + \gamma.
\end{aligned} \tag{9}$$

The posterior density in (9) is a mixture of densities of the same form as in (8). The weight w_t is the posterior probability that t out of $n - \gamma$ women recorded as Untalented were correctly recorded. And, the posterior density is a mixture of $n - \gamma + 1$ possible posterior densities that would result under perfect knowledge of the exact number of misclassifications (i.e., under perfect knowledge of $n - \gamma - t$). The marginal posterior densities for θ and λ can be obtained as

$$\begin{gathered}
f(\theta|\gamma, n, t) = \sum_{t=0}^{n-\gamma} w_t f_\beta(\theta|a_\theta^*, b_\theta^*)\\
\text{and}\\
f(\lambda|\gamma, n) = \sum_{t=0}^{n-\gamma} w_t f_\beta(\lambda|a_\lambda^*, b_\lambda^*).
\end{gathered} \tag{10}$$

To get a feel for these results, let's consider Recruitment. We are uncertain about the actual proportion of Talented women, but suspect that $\theta \sim f_\beta(\theta|7, 3)$. This implies that *a priori*, $\mathrm{E}(\theta) = 0.7$, but the distribution is fairly spread out and admits all values of θ between 0 and 1. At the same time, we suspect that there is a consequential chance of women being incorrectly recorded as Untalented in any data that we might see. Suppose that our uncertainty about the one-sided misclassification rate (false negative) is best represented by $\lambda \sim f_\beta(\lambda|3, 7)$. Then, $\mathrm{E}(\lambda) = 0.3$, and as in the case of θ, the prior distribution for λ is also quite spread out, admitting values close to zero and as high as 0.7. These marginal prior densities for θ and λ are shown in Figure 6 below.

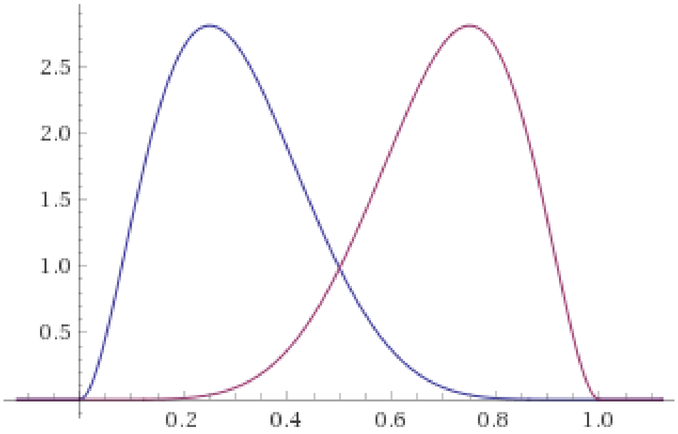

Figure 6: Prior distributions for $\theta \sim \text{Beta}(7, 3)$ (red curve) and $\lambda \sim \text{Beta}(3, 7)$ (blue curve).

Now suppose we see data on 100,000 women of whom 30% are recorded as Talented, as in Recruitment. Given our model, one thing is immediately clear, that the number of women who are actually Untalented is inflated in the recorded data. Our conjecture, built into the model, is that this is the result of historical patterns of discrimination and implicit biases leading to women being perceived as on average less Talented. But, we are uncertain as to what extent. Using our model, we can find the posterior distributions for θ and λ. As can be seen from the above expressions, the analytical calculations require computing combinatorial terms with large values. Hence, we use stochastic simulation to approximate the posterior marginal densities. For example, Figure 7 illustrates the marginal posterior densities for θ and λ that we obtained given the data ($\gamma = 30,000$ and $n = 100,000$).

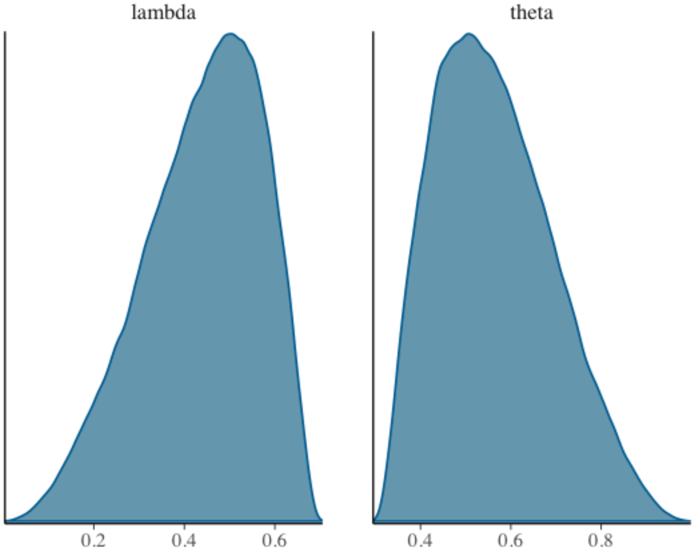

Figure 7: Posterior distributions for θ and λ for $n = 10000$.

We obtained the posterior densities and moments using a Markov Chain Monte Carlo algorithm known as the Gibbs Sampler implemented in a Bayesian MCMC inference program called JAGS (Plummer, 2003). The Gibbs sampler was first introduced by Geman and Geman (1984) in the context of image processing.

Recall that the marginal posterior mean for θ is our predictive probability that Alice in our Recruitment example is Talented, using Huttegger's Generalized Rule of Succession. The posterior mean for θ that we obtained using this simulation is 0.55. Note that despite the very large sample suggesting that only 30% of the women are Talented, our posterior mean for the proportion of women who are Talented is 0.55, i.e., 55% of the women. In other words, we avoid the Sad Conclusion! Our guess would be that Alice is slightly more likely to be Talented than Untalented. Note that this result is obtained without any considerations of epistemic risk (as in the JKB model) in our prior for θ, but simply by "epistemic-risk-neutral" representation of our uncertainty about θ and the misclassification rate λ.

To get further insight concerning this model, it is useful to consider the entire posterior density for θ which provides the full representation of uncertainty about θ, and compare it to a noise-free model. Ignoring misclassifications, i.e., assuming $\lambda = 0$, the prior $f(\theta) = f_\beta(\theta|a_\theta = 7, b_\theta = 3)$ would be revised to the posterior $f(\theta|\gamma, n) = f_\beta(\theta|a_\theta + \gamma = 30007, b_\theta + n - \gamma = 70003)$. Note that the prior mean of θ is 0.7. The noise-free posterior density has a mean of 0.3, with a standard deviation of 0.001, placing almost the entire probability mass at $\theta = 0.3$ and completely overwhelming the prior on θ. On the other hand, in our model with noise, the posterior mean of θ is almost 0.5, with a posterior standard deviation of 0.129 which is 129 times larger than in the noise-free case. This is because consideration of all the possible misclassifications that could have occurred in the sample leads to much greater uncertainty about θ. In fact, a noise-free sample of merely **four** women would lead to the same posterior mean and standard deviation as in our model with noise. Table 1 shows the equivalent noise-free sample size that would be needed to obtain the same posterior mean and same posterior standard deviation for θ as in our model with noise, with $\theta \sim f_\beta(\theta|7,3)$ and $\lambda \sim f_\beta(\lambda|3,7)$, as n increases from 10 to infinity. In the Appendix, we show the analytical results on the noise-free equivalent sample size in case of an infinitely large sample (Theorem 1 in the Appendix enables us to compute the last row, with $n \to \infty$).

Sample size (n)	Equivalent noise-free sample size for θ (n^*_θ)
10	1.4
100	3.2
$1,000$	3.6
$10,000$	3.7
$100,000$	3.7
$n \to \infty$	3.8

Table 1: Actual and equivalent noise-free sample sizes

Note that in our example above, the equivalent noise free sample size has an upper bound of 3.8. This reveals a drastic loss of information in the sample. This is because, as mentioned before, the likelihood function is unable to distinguish between numerous disparate explanations of the observed data (i.e., an infinite combination of θ and λ lead to the same likelihood). Prior distributions on θ and λ help discriminate between (θ, λ) pairs with the same likelihood. Hence, with diffuse priors on θ and λ, the resulting marginal posterior distributions for θ and λ are almost flat over [0,1], as like the likelihood function even the priors do not discriminate between the (θ, λ) pairs. Gaba and Winkler (1992), for example, show how the loss of information (i.e., the equivalent noise-free sample size) is impacted by different prior distributions for λ. The equivalent noise-free sample size relative to the actual sample size remains disproportionately low even for tight distributions for λ or low expected values of λ. The point is that we do not have to bury our head in the sand, or overload our prior in an unrealistic manner. Rather, simply by using the information we have regarding gender disparities in unemployment, historical practices of discrimination, etc., and accepting the possibility of observed data coming from a noisy process (i.e., the possibility of Talented women being classified as Untalented in cases like Recruitment), one will avoid normative conflicts regardless of sample size! The potential for such misclassifications is a serious risk in contexts like this.

7 Concluding Remarks

We have shown in this project that one can avoid normative conflicts even with arbitrarily large samples provided we accept that the data are generated from a noisy process. In response to Gendler's concern, then, it is rarely the case that a perfectly rational decision maker should manifest different behaviors toward members of different races, even if there exists a differential crime rate, provided, of course, that the difference is occurring at least in part due to police bias, racial profiling, bigotry, over-prosecution of minorities, etc. And indeed, we suspect it is factors like this that in fact account for differences in crime rates across race. A similar conclusion can be made for performance differences in academia, finance, etc. across many different vulnerable groups, those characterized by factors such as race, ethnicity, and gender. Thus, in most real life cases normative conflicts are avoidable. But the requirement of noise does mean that we cannot guarantee it will be avoided. The absence of such a guarantee is a virtue, rather than a vice, however. It means that when there are true underlying population differences – those that should not be ignored – our model will enable us to learn them.

Appendix

In this Appendix we derive the posterior distribution for θ when $n \to \infty$ with the proportion of women who are deemed Talented, $\varphi_0 = \gamma/n$, being held constant. We will exploit the fact that as $n \to \infty$, the likelihood (7) converges to the Dirac delta function $\delta\left(\theta\left(1-\lambda\right)-\varphi_0\right)$.

Theorem 1. Let the prior distribution for (θ, λ) be a product of independent beta distributions $f_\beta(\theta|a_\theta, b_\theta)$ and $f_\beta(\lambda|a_\lambda, b_\lambda)$, and let the likelihood be the Dirac delta function $\delta\left(\theta\left(1-\lambda\right)-\varphi_0\right)$. Then, up to normalization, the marginal posterior pdf for θ is

$$f\left(\theta|\varphi_0\right) \propto \begin{cases} \theta^{a_\theta - a_\lambda - b_\lambda}\left(1-\theta\right)^{b_\theta - 1}\left(\theta - \varphi_0\right)^{a_\lambda - 1} \text{ if } \theta \geq \varphi_0, \\ 0 \text{ if } \theta < \varphi_0. \end{cases}$$

Proof. When $n \to \infty$, the joint posterior pdf is

$$f(\theta, \lambda|\varphi_0) \propto f_\beta(\theta|a_\theta, b_\theta) f_\beta(\lambda|a_\lambda, b_\lambda) \delta\left(\theta\left(1-\lambda\right)-\varphi_0\right),$$

and the marginal posterior for θ is

$$f\left(\theta|\varphi_0\right) \propto f_\beta(\theta|a_\theta, b_\theta)\left(\int_0^1 f_\beta(\lambda|a_\lambda, b_\lambda)\delta\left(\theta\left(1-\lambda\right)-\varphi_0\right)d\lambda\right)$$

.

Denote $\varphi = \theta\left(1-\lambda\right)$; then $\lambda = 1-\varphi/\theta$, $d\lambda = -\frac{1}{\theta}d\varphi$, and

$$\begin{aligned}&\int_0^1 f_\beta(\lambda|a_\lambda, b_\lambda)\delta\left(\theta\left(1-\lambda\right)-\varphi_0\right)d\lambda\\ &= \int_0^\theta f_\beta(1-\varphi/\theta|a_\lambda, b_\lambda)\delta\left(\varphi-\varphi_0\right)\frac{1}{\theta}d\varphi\\ &= \begin{cases}\frac{1}{\theta}f_\beta\left(1-\varphi_0/\theta|a_\lambda, b_\lambda\right) \text{ if } \theta \geq \varphi_0,\\ 0 \text{ if } \theta < \varphi_0.\end{cases}\end{aligned}$$

Therefore,

$$f\left(\theta|\varphi_0\right) \propto \begin{cases}\frac{1}{\theta}f_\beta(\theta|a_\theta, b_\theta)f_\beta\left(1-\varphi_0/\theta|a_\lambda, b_\lambda\right) \text{ if } \theta \geq \varphi_0,\\ 0 \text{ if } \theta < \varphi_0.\end{cases}$$

In turn, $f_\beta(\theta|a_\theta, b_\theta) \propto \theta^{a_\theta-1}\left(1-\theta\right)^{b_\theta-1}$ and $f_\beta(\lambda|a_\lambda, b_\lambda) \propto \theta^{a_\lambda-1}\left(1-\theta\right)^{b_\lambda-1}$, so

$$\begin{aligned}f\left(\theta|\varphi_0\right) &\propto \begin{cases}\frac{1}{\theta}\theta^{a_\theta-1}\left(1-\theta\right)^{b_\theta-1}\left(1-\varphi_0/\theta\right)^{a_\lambda-1}\left(\varphi_0/\theta\right)^{b_\lambda-1} \text{ if } \theta \geq \varphi_0,\\ 0 \text{ if } \theta < \varphi_0\end{cases}\\ &\propto \begin{cases}\theta^{a_\theta-a_\lambda-b_\lambda}\left(1-\theta\right)^{b_\theta-1}\left(\theta-\varphi_0\right)^{a_\lambda-1} \text{ if } \theta \geq \varphi_0,\\ 0 \text{ if } \theta < \varphi_0.\end{cases}\end{aligned}$$

□

The marginal posterior pdf for λ is derived similarly, and is given by

$$f\left(\lambda|\varphi_0\right) \propto \begin{cases}\lambda^{a_\lambda-1}\left(1-\lambda\right)^{b_\lambda-a_\theta-b_\theta}\left(1-\varphi_0-\lambda\right)^{b_\theta-1} \text{ if } \lambda \leq 1-\varphi_0,\\ 0 \text{ if } \lambda > 1-\varphi_0.\end{cases}$$

To compute the equivalent sample size as in Table 1, we first find parameters $a_\theta^*$and b_θ^* of the beta distribution that match the first two moments of $f\left(\theta|\varphi_0\right)$. Then the equivalent sample size equals $n_\theta^* = a_\theta^* + b_\theta^* - a_\theta - b_\theta$. From the equations $E\left(\theta|\varphi_0\right) = \frac{a_\theta^*}{a_\theta^*+b_\theta^*}$ and $V\left(\theta|\varphi_0\right) = \frac{a_\theta^* b_\theta^*}{\left(a_\theta^*+b_\theta^*\right)^2\left(a_\theta^*+b_\theta^*+1\right)}$ we find $a_\theta^* + b_\theta^* = \frac{E(\theta|\varphi_0)(1-E(\theta|\varphi_0))}{V(\theta|\varphi_0)} - 1$. In particular, for $a_\theta = 7$, $b_\theta = 3$, $a_\lambda = 3$, $b_\lambda = 7$, and $\varphi_0 = 0.3$ we get $n_\theta^* = 3.8$.

References

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (May 23, 2016). Machine Bias.

Babic, B. (2019). A Theory of Epistemic Risk. *Philosophy of Science 86*(3), 522–550.

Basu, R. (2018). The Specter of Normative Conflict. In E. Beeghly and A. Madva (Eds.), *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*. London: Routledge.

Basu, R. and M. Schroeder (2018). Can Beliefs Wrong? In *Philosophical Topics (Special Issue)*. Forthcoming.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review 78*(1), 1–3.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Gaba, A. (1993). Inferences with an Unknown Noise Level in a Bernoulli Process. *Management Science 39*(10), 1179–1197.

Gaba, A. and R. L. Winkler (1992). Implications of Errors in Survey Data: A Bayesian Model. *Management Science 38*(7), 913–925.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis* (3rd ed.). New York: CRC Press (Taylor & Francis).

Geman, S. and D. Geman (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machien Intelligence 6*, 721–724.

Gendler, T. S. (2011). On the Epistemic Costs of Implicit Bias. *Philosophical Studies 156*(1), 33–63.

Greaves, H. and D. Wallace (2006). Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility. *Mind 115*(459), 607–632.

Huttegger, S. M. (2013). In Defense of Reflection. *Philosophy of Science 80*(3), 413–433.

Huttegger, S. M. (2017). *The Probabilistic Foundations of Rational Learning*. Cambridge: Cambridge University Press.

Johnson, W. (1924). *Logic, Part III. The Logical Foundation of Science.* Cambridge: Cambridge University Press.

Johnson-King, Z. and B. Babic (2019). Moral Obligation and Epistemic Risk. *Oxford Studies in Normative Ethics 10.*

Joyce, J. M. (1998). A Nonpragmatic Vindication of Probabilism. *Philosophy of Science 65*, 575–603.

Joyce, J. M. (2009). Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In F. Huber and C. Shmidt-Petri (Eds.), *Degrees of Belief*, pp. 263–300. Springer.

Kleinberg, J., S. Mullainathan, and M. Raghavan (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807.*

Lindley, D. V. and L. Phillips (1976). Inference for a Bernoulli Process (A Bayesian View). *The American Statistician 30*(3), 112–119.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence.* Oxford: Oxford University Press.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling.

Savage, L. J. (1971). Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association 66*(336), pp. 783–801.

Winkler, R. L. and A. Gaba (1990). Inference with Imperfect Sampling from a Bernoulli Process. In S. P. S Geisser, J.S. Hodges and A. Zellner (Eds.), *Bayesian and Likelihood Methods in Statistics and Econometrics*, pp. 303–317. North-Holland.

Zabell, S. L. (2005). *Symmetry and its Discontents: Essays on the History of Inductive Probability.* Cambridge: Cambridge University Press.